National Défense
Defence nationale

A Handbook on the Theory and Methods of Differential Item Functioning (DIF)

LOGISTIC REGRESSION MODELING AS A UNITARY FRAMEWORK FOR BINARY AND LIKERT-TYPE (ORDINAL) ITEM SCORES

Bruno D. Zumbo, Ph.D.
Professor of Psychology & of Mathematics
University of Northern British Columbia

April 1999

Approved by:
W.R. Wild
Lieutenant-Colonel
Director

Directorate of Human Resources Research and Evaluation
National Defense Headquarters
Ottawa, Canada
K1A 0K2

Canada

Citation format:
Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Precis / Summary

That a test not be biased is an important consideration in the selection and use of any psychological test. That is, it is essential that a test is fair to all applicants, and is not biased against a segment of the applicant population. Bias can result in systematic errors that distort the inferences made in selection and classification. In many cases, test items are biased due to the fact that they contain sources of difficulty that are irrelevant or extraneous to the construct being measured, and these extraneous or irrelevant factors affect performance. Perhaps the item is tapping a secondary factor or factors over-and-above the one of interest. This issue, known as test bias, has been the subject of a great deal of recent research, and a technique called Differential Item Functioning (DIF) analysis has become the new standard in psychometric bias analysis. The purpose of this handbook is to provide a perspective and foundation for DIF, a review and recommendation of a family of statistical techniques (i.e., logistic regression) to conduct DIF analyses, and a series of examples to motivate its use in practice. DIF statistical techniques are based on the principle that if different groups of test-takers (e.g., males and females) have roughly the same level of something (e.g., knowledge), then they should perform similarly on individual test items regardless of group membership. In their essence, all DIF techniques match test takers from different groups according to their total test scores and then investigate how the different groups performed on individual test items to determine whether the test items are creating problems for a particular group.

Introduction

The Canadian Forces currently uses psychological tests in the selection and assessment of its personnel. In addition, other psychological tests are in various stages of evaluation for introduction into the Canadian Forces selection system. A consideration in the use of any psychological or educational test for selection is that the test is fair to all applicants, and is not biased against a segment of the applicant population. This issue, known as test bias, has been the subject of a great deal of recent research, and a technique called Differential Item Functioning (DIF) analysis has become the new standard in test bias analysis.

As active users of psychological tests, it is necessary for researchers, policy makers, and personnel selection officers involved in the evaluation of tests to be conversant with the current thinking in test bias analysis. With an eye toward this purpose, this handbook will provide:

- the theoretical background for the DIF techniques, including a foundation in current psychometric validity theory and a definition of the relevant terms;
- a step-by-step guide for performing a DIF analysis; and
- annotated examples of the performance of the DIF techniques.

Before proceeding it is important to define two types of scores that can arise from tests and measures. The two most commonly used *scoring formats* for tests and measures are binary and ordinal. Binary scores are also referred to as dichotomous item responses and ordinal item responses are also referred to as graded response, Likert, Likert-type, or polytomous[1]. The ordinal formats are commonly found in personality, social, or attitudinal measures. For simplicity and consistency with the statistical science literature, throughout this handbook these various terms denoting ordered multicategory scores will be referred to as "ordinal".

It is important to note that it is not the question format that is important here but the scoring format. Items that are scored in a binary format are either: (a) items (e.g., multiple choice) that are scored correct/incorrect in aptitude or achievement tests, or (b) items (e.g., true/false) that are dichotomously scored according to a scoring key in a personality scale. Items that are scored according to an ordinal scale may include Likert-type scales such as a 5-point strongly agree to strongly disagree scale on a personality or attitude measure.

This handbook is organized into five sections. In the first section of the handbook (Current Conceptions of Validity Theory With an Eye to Item Bias) I provide a broad perspective on bias analysis linking it to current conceptions of validity theory in psychological measurement. In the second section (Concepts and Definitions of Bias, Item Bias, and DIF: Relevant terminology, concepts, and policy issues) I introduce some basic concepts and definitions of terms found in the scientific and practice literature on

[1] Note that in this context we are using polytomous to imply ordered responses and not simply multicategory nominal responses.

bias. In this section I also briefly discuss some policy issues that need to be addressed in establishing a procedure for DIF analysis. The third section (Statistical Methods For Item Analysis) acts as a transition section from foundations and principles to statistical methods for bias analysis, of which DIF is an example. In the fourth section (A Statistical Method For DIF: Logistic Regression) I introduce logistic regression analysis as the statistical method of choice for DIF analysis. I describe a well known formulation of logistic regression for binary scored items (e.g., correct/incorrect or true/false) and I introduce a new methodology for the analysis of ordinal responses. This methodology is a natural extension of the logistic regression method for binary items. For both binary and ordinal logistic regression, new measures are introduced and applied to help one get a sense of the magnitude of DIF. That is, for DIF to be useful it is not sufficient to simply have a statistical test of DIF but one also needs a measure of the magnitude of the DIF effect. And finally, the handbook concludes with a section containing examples from both binary and ordinal item responses. This last section will focus on statistical computer programming in SPSS.

Current Conceptions Of Validity Theory With An Eye To Item Bias

Technological and theoretical changes over the past few decades have altered the way we think about test validity. This section will briefly address the major issues and changes in test validity with an eye toward bias analysis.

Evaluating the measures: Validity and Scale Development

The concept, method, and process of validation is central to evaluating measures, for without validation, any inferences made from a measure are meaningless. Throughout this presentation, the terms measure, observation, score, test, index, indicator, and scale will be used interchangeably and in their broadest senses to mean any coding or summarization of observed phenomenon.

This previous paragraph highlights two central features in contemporary thinking in validation.

- First, it is not a measure that is being validated but rather the inferences one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from a scale may appear at first blush subtle but in fact it has significant implications for the field of assessment. I will discuss these implications later in this section when I describe the current validation theory and process.
- The second central feature in the above paragraph is the clear statement that all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an “objective” human resource/health/social indicator such as number of sick days, or a more psychological measure such as a depression scale or a measure of life satisfaction and well-being, one must be concerned with the validity of the inferences.

In recent years, there has been a resurgence of thinking about validity in the field of testing and assessment. This resurgence has been partly motivated by the desire to expand the traditional views of validity to incorporate developments in qualitative methodology and in concerns about the consequences of decisions made as a result of the assessment process.

For a review of current issues in validity I recommend a recent paper by Hubley and Zumbo (1996) and the edited volume by Zumbo (1998).

Let me now contrast the traditional and more current views of validity.

Traditional View of Validity.

The traditional view of validity focuses on:

- whether a scale is measuring what we think it is,
- reliability as a necessary but not sufficient condition for validity
- validity as a property of the measurement tool,
- validity as defined by a set of statistical methodologies, such as correlation with a gold-standard,

- a measure is either valid or invalid, and
- various types of validity -- usually four -- and as Hubley and Zumbo (1996) state, in practice the test user or researcher assumes that they only need to demonstrate one of the four types to have demonstrated validity.

Table 1 describes the traditional view of validity.

Table 1.
The traditional categories of validity (based on Hubley and Zumbo, 1996, p. 209)

Type of Validity	**What do we do to show this type of validity?**
Content	Ask experts if the items (or behaviors) tap the construct of interest.
Criterion-related:	
A. Concurrent	Select a criterion and correlate the measure of interest with the criterion obtained in the present
B. Predictive	Select a criterion and correlate the measure of interest with the criterion obtained in the future
Construct	Can be done several different ways. Some common ones are (a) correlate to a “gold standard”, (b) a statistical technique called factor analysis, (c) convergent and discriminant validity

The process of validation then simply becomes picking the most suitable strategy from Table 1 and conducting the statistical analyses. For example, if we were conducting a human resource study on work environment and had developed a new measure of job satisfaction a common strategy would be to conduct a study to see how well the new measure correlates with some gold standard (e.g., the quality of work life scale) and if the correlation is sufficiently large then we would be satisfied that the measure is valid. This correlation with the gold standard is commonly referred to as a validity coefficient. Ironically, of course, the gold standard may have been developed and validated with some other gold standard.

The current view of validity.

It is important to note that there is, of course, nothing inherently wrong with the traditional view of validity. The purpose of the more current view of validity is to expand the conceptual framework of the traditional view of validity.

To help us get a better working knowledge of the more current conceptions of validity let me restate the traditional features listed above in the following way:

- construct validity is the central most important feature of validity and one must show construct validity;

- there is debate as to whether reliability is a necessary but not sufficient condition for validity; my view is that this issue is better cast as one of measurement precision so that one strives to have as little measurement error as possible in ones inferences;
- validity is no longer a property of the measurement tool but rather of the inferences made from that tool;
- the validity conclusion is on a continuum and not simply declared as valid or invalid;
- validity is no longer defined by a set of statistical methodologies, such as correlation with a gold-standard but rather by an elaborated theory and supporting methods;
- consequences of *test decisions* and *use* are an essential part of validation; and
- there are no longer various types of validity so that it is no longer acceptable in common practice that the test user or researcher assumes that he/she only needs to demonstrate one of the four types to have validity.

As an example to help motivate our understanding of the more current thinking in validity let us consider the example of validating a job satisfaction measure. First, one is interested in gathering evidence supporting the trustworthiness that the scale actually measures job satisfaction and not some other related construct (such as coping with emotional stress, mental distress, general life dissatisfaction). To do that, one might consider:

- correlations with other theoretically related (i.e., convergent) and unrelated (i.e., discriminant) constructs,
- factor analysis of the scale,
- focus groups of target samples/groups of men and women, and expected age, or other differences to explore the consequential basis of validity.

To continue with the example, next, one would want to amass information about the value implications of the construct label itself, the broader theory of job satisfaction in women, and even broader ideologies about work life. For example, what is the implication of labeling a series of behaviors or responses to items as “high job satisfaction” and what does this mean for human resource policy, “do all people have to be happy and satisfied with their job, or is it sufficient to simply be able at their job tasks?."

And, if we want to use the measure in decision-making (or, in fact, simply use it in research) we need to conduct research to make sure that we do not have bias in our measures. Where our value statements come in here is that we need to have organizationally and socially relevant comparison groups (e.g., gender or minority status).

As you can see from the simple example of job satisfaction, the traditional view of validity (as it is usually used in research; that is, either a correlation or just a factor analysis) is quite meager compared to the more comprehensive current approach. Furthermore, the traditional view of validity is not tossed out but rather built on and expanded. Basically, the current view of validity makes validation a core issue that is not resolved by simply computing a correlation with another measure (or even a factor analysis of the items).

Another implication of the current view of validity is that we now need explicit statistical studies examining bias and concept-focused and policy studies of value implications for research use and decision making. Much of this has gone on in either informal or unsynthesized ways but now we need to bring this all together to address the issue of the inferences we make from measures. In a recent paper in my 1998 edited volume, Messick clarifies the issue of consequences. He states that it is not the obvious misuses of measures that is the issue but rather that we need to think about the unanticipated (negative and positive) consequences of the legitimate use and/or interpretation for decision making from measures that can be traced back to test invalidity -- such as construct under-representation and/or construct irrelevant variance. Item bias studies are examples of the sort of questions that need to be asked. That is, the validation process begins at the construct definition stage before items are written or a measure is selected, continues through item analysis (even if one is adopting a known measure), and needs to continue when the measure is in use.

In conclusion, it should be noted that what I have described as the traditional view of validity is, in fact, somewhat of an extreme portrayal (although you will see lots of this extreme behavior in everyday practice and in many human resource and scientific journals). Although modern validity has been around in a somewhat complete form at least since Messick's 1988 chapter on the topic, the practice of validity has reflected *parts* of it for some time. The modern approach to validity demands long-term commitment of resources and personnel both which are often in short supply. The reason why one needs such a commitment is because the validation process is never entirely complete. One is always doing some study of consequences of test use, measurement error issues, and the context of use does evolve. This is not as daunting as it sounds and in fact probably describes what most testing and human resources organizations are doing already -- at least to some extent.

Concepts And Definitions Of Bias, Item Bias, And DIF: Relevant Terminology, Concepts, And Policy Issues

Now that I have laid the foundation and motivation for a bias analysis, let's turn to the relevant terminology, concepts, and policy issues. The definitions presented here are taken from Zumbo and Hubley (1998 a) and are based on Camilli and Shepard (1994) and Clauser and Mazor (1998).

Item analysis: A set of statistical techniques to examine the performance of individual items. This is important when developing a test or when adopting a known measure.

Item impact: Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item.

DIF: DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure.

Item bias: Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias.

Adverse Impact: Adverse impact is a legal term describing the situation in which group differences in test performance result in disproportionate examinee selection or related decisions (e.g., promotion). This is *not* evidence for test bias.

Item impact and item bias differ in terms of whether group differences are based on relevant or irrelevant characteristics (respectively) of the test. DIF requires that members of the two groups be matched on the relevant underlying ability before determining whether members of the two groups differ in their probability for success. And finally, adverse impact simply describes disproportionate workplace decisions based on test performance.

As mentioned above, DIF is a necessary, but not sufficient, condition for item bias. Thus, if DIF is *not* apparent for an item, then no item bias is present. However, if DIF is apparent, then its presence is *not* sufficient to declare item bias; rather, one would have to apply follow-up item bias analyses (e.g., content analysis, empirical evaluation) to determine the presence of item bias.

As Clauser and Mazor (1998) remind us, it is important to distinguish both item bias and DIF from inappropriate item content or framing that is potentially offensive. Most testing and human resource organizations deal with inappropriate item content by formal (or informal) screening panels to eliminate items with offensive language. What is important to keep in mind is that if an item is offensive to everyone, it is not going to be detected as biased -- by the simple fact that lack of bias means that no one group is affected but rather both (or all) groups are equally affected.

Two approaches to examining potential measurement bias are: (a) judgmental, and (b) statistical. Judgement methods rely solely on one or more expert judges' opinions to select potentially biased items. Clearly, this is an impressionistic methodology. Instead of the sole reliance on expert (i.e., content area) judges, I recommend that in a high-stakes context faced by human resource organizations one rely on statistical techniques for investigating potential bias because of this method's defensibility. The statistical technique then flags potentially biased items.

One can study potential measurement bias by investigating *external* relationships or *internal* relationships. The former commonly focuses on predictive studies investigating criterion-scale relationships such as those in some personnel selection studies. External evidence of measurement bias is suggested when the relationship between the scores and criterion is different for the various groups. Internal evidence comes from the internal relationships of the items to each other. DIF is clearly a matter of internal item relationships.

Note that through all of this discussion I am focusing on the use of composite scores (i.e., scale total scores) as indicators of a latent (unobservable) variable. The expressions *latent variable* and *unobserved variable* are used interchangeably in this discussion. The latent or unobserved variable is the variable we are trying to get at with our indicators (such as items). In most, if not all, of our analyses and discussion the latent variable represents a *quantity*. That is, respondents have an amount of this latent variable -- for example, an amount of intelligence, verbal ability, spatial ability, sociability, etc.. I will come back to this latent variable (also called a continuum of variation by which I mean a continuum upon which individuals vary) later and I will discuss how its inclusion in item and test analysis has revolutionized psychometrics.

Policy Issues and Issues of Implementing a DIF Screening Strategy

As Clauser and Mazor (1998) remind us, the entire domain of item bias is really about policy. In fact, by even considering a bias analysis one is already in the domain of policy. Some organizations have bias analysis legislated whereas others take it on as part of the day-to-day validation process. If bias is being “legislated” from an outside body, this legislation will help you determine the answer to the following policy matters.

I see five operational policy matters of essential concern:

1. There are a lot of different sub-groups one could contrast. You need to be clear as to which ones are of personal and moral focus. The standard comparisons are based on gender, race, sub-culture, or language.

2. You need to discuss how much DIF do you need to see before you are willing to consider the item as displaying DIF. In most cases, it is not sufficient to simply rely on the answer that all statistically significant items are displaying DIF because statistical power plays havec on your ability to detect effects. Please see Zumbo and Hubley (1998 b) in the *Journal of the Royal Statistical Society* for further detailed

statistical discussion of this matter. In essence, how much DIF does one need to see before one puts the item under review or study.

3. Should an item only be sent for review if it is identified as favoring the reference group (e.g., the majority group)? Or should an item be sent for review irrespective of whether it favors the reference or focal group?

4. The timing of the DIF analysis is also important. Let me consider two scenarios (a) you are using a ready-made test; or (b) you are developing your own new or modified measure. First of all, in either case, DIF analyses are necessary. In the first scenario where you have a ready-made test that you are adopting for use *and* there is pilot testing planned with a large enough sample (I address this later in this document), DIF analyses are performed at the pilot testing stage. When you do not have a pilot study planned or you do not have a large enough pilot sample then DIF analyses are conducted before final scoring is done and therefore before scores are reported. In the second scenario where you are developing a new test, DIF analyses should be conducted at pilot testing and certainly before any norms or cut-off scores are established.

5. What does one do when one concludes that an item is demonstrating DIF? Does one immediately dispense with the item (I won't recommend this because your domain being tapped will quickly become too limited) or does one put an item "on ice" until one sends it to content experts and for further validation studies? Part of the answer to this question has to do with the seriousness of a measurement decision (or more importantly the seriousness of making an error).

Statistical Methods For Item Analysis

In this section I will discuss further the continuum of variation of a latent variable. The continuum of variation is an essential part of modern test theory and it not only helps define DIF but it also helps us interpret DIF results. This section highlights for us that DIF is an item analysis methodology.

In the previous section I made reference to the point that a latent variable represents a quantity. That is, respondents have an amount of this latent variable. For example, an amount of intelligence, verbal ability, spatial ability, sociability, etc., and that it is a continuum along which individuals vary. That is, different people may have a different amount of the latent variable we are trying to tap with our items. The term continuum of variation is far less politically laden than the term “latent trait.” This concept of a continuum of variation is fundamental to modern test theory.

Modern test theory, in general, according to Hambleton, Swaminathan, and Rogers (1991), is based on two postulates:

1. that the performance of an examinee on a test item can be explained or predicted from a set of factors traditionally called "traits, latent traits, or abilities" -- which we refer to as a continuum of variation; and
2. that the relationship between examinees' item performance and the continuum of variation underlying performance on that item can be described as an item characteristic curve (ICC) or item response function (IRF).

A parametric ICC is the monotonically increasing function to which items are fit in most item response models. Historically, the function was the normal ogive function but was later replaced with the more tractable logistic function. Parametric ICCs vary in terms of their position on the X-axis, their slope, and their intercept with the Y-axis. The X-axis is the latent variable and the Y-axis is the probability of getting the item correct (or endorsing the item).

The following terms will help you understand DIF. In the following table, I adapt terminology discussed in Zumbo, Pope, Watson, & Hubley (1997):

Table 2.
Interpretation of ICC Properties for Cognitive and Personality Measures

ICC Property	**Cognitive, Aptitude, Achievement, or Knowledge Test**	**Personality, Social, or Attitude Measures**
Position along the X-axis (commonly called the b-parameter in IRT)	Item difficulty Amount of a latent variable needed to get an item right	Threshold Amount of a latent variable needed to endorse the item
Slope (commonly called the a-parameter in IRT)	Item discrimination. A flat ICC does not differentiate among test-takers	Item discrimination. A flat ICC does not differentiate among test-takers
Y-Intercept (commonly called the c-parameter in IRT)	Guessing	The likelihood of indiscriminate responding or social desirable responses

All of these parts of an ICC provide detailed information on various aspects of the item. Figure 1 gives an example of a parametric ICC. The horizontal axis is the continuum of variation for the latent variable. The scale of the latent variable is in z-scores. The item depicted in Figure 1 has a discrimination (i.e., slope) of 2.0, difficulty (i.e., threshold) of 0.60, and guessing parameter of 0.10.

Note that the item discrimination parameter determines how rapidly the curve rises from its lowest value of *c*, in this case 0.10, to 1.0. Note that if the curve is relatively flat then the item does not discriminate among individuals with high, moderate, or low total scores on the measure. Item discrimination values of 1.0 or greater are considered very good. Finally, note that the threshold parameter is the latent variable value on the continuum of variation at which the curve is midway between the lowest value, *c*, and 1.0, and therefore for achievement or aptitude measures, is a marker of the item difficulty. Items with difficulty values less than -1.0 indicate a fairly easy item whereas items with difficulty greater than 1.0 indicate rather difficult items.

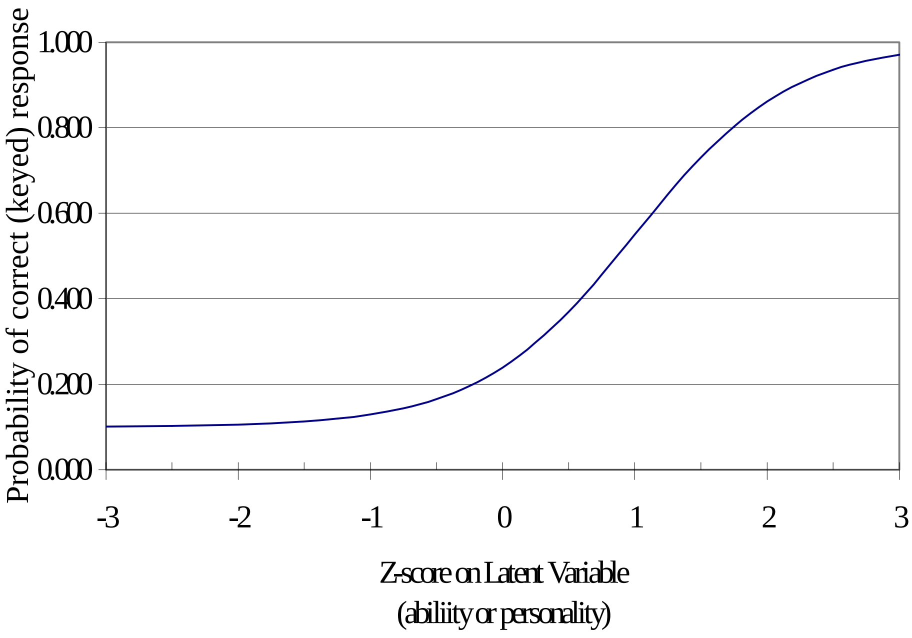


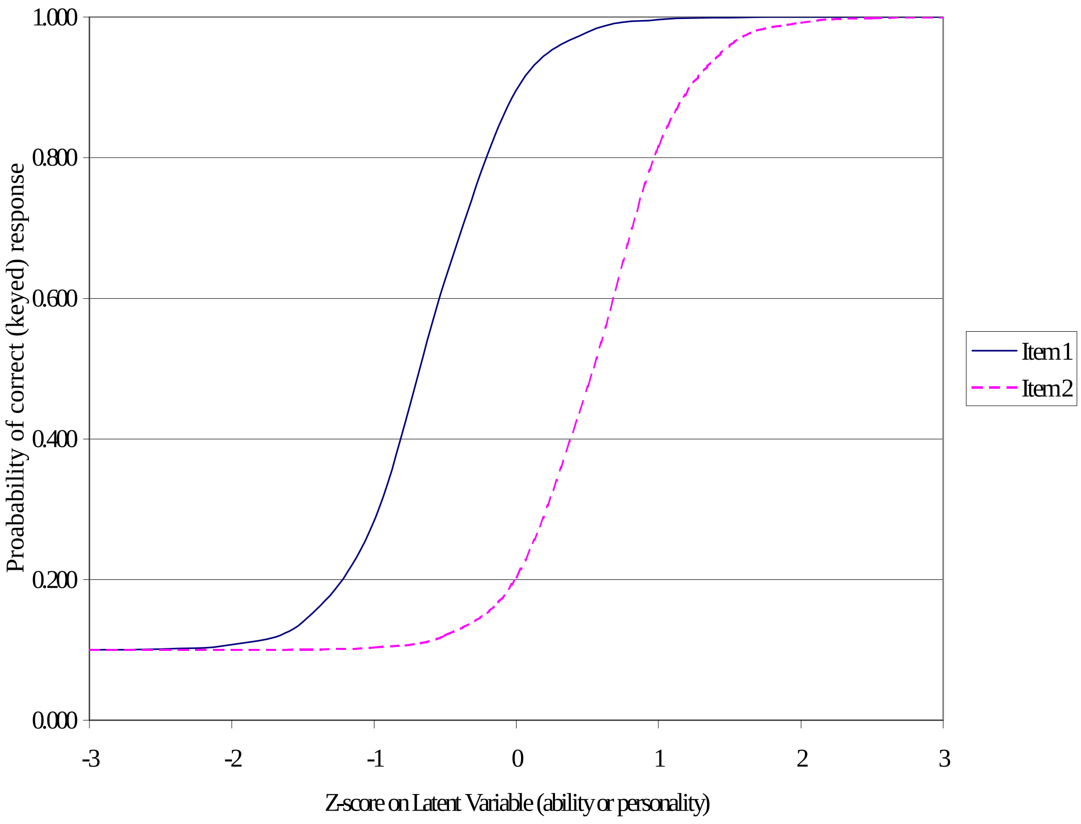
Figure 1. A logistic item characteristic curve

The ICCs shown in Figure 2 portray two items with equal slope (i.e., equal discrimination among respondents) but different placements on the continuum of variation. One would need to have more of the latent variable to endorse the item depicted by the dashed line than by the solid line. The dashed line is further to the right on the continuum. The dashed line, item 2, is thus considered more difficult.

In summary then, the ICC depicts the non-linear regression of the probability of the correct (keyed) response onto the continuum of variation conceptualized as the latent variable of interest. If we take an aptitude test as an example,

- the y-intercept is the likelihood of someone of very low aptitude getting the item correct,
- the slope of the curve indicates how well the item discriminates among levels of ability (a flat curve, for example, would mean that irrespective of the level of ability, individuals have the same likelihood of getting an item correct),
- the threshold indicates that value on the continuum of variation (the X-axis) at which the probability of getting the item correct is midway between the lowest value, *c*, and 1.0. For items for which the guessing parameter is zero, this midway value of course represents the value of the latent variable at which the likelihood of a correct response is greater than 0.50.

Figure 2. Two item characteristic curves with different positions on the continuum of variation



The continuum of variation has revolutionized psychometrics because traditional classical test theory methods are summary omnibus statistics that are, in essence, averages across the continuum of variation. For example, item total correlations (a traditional discrimination statistic) or reliability coefficients are one number irrespective of the level of individual variation. That is, the coefficient alpha is thought of as the same number irrespective of whether the respondent scored 3 standard deviations below, at, or 3 standard deviations above, the mean.

Therefore, these summary measures describe the sample as whole, ignoring how the psychometric properties of the scale may vary as a function of variation within the sample. Modern test theory builds on classical test methods and takes into account this sample variation and continuum of variation.

One natural psychometric technique that has arisen in this tide of modern psychometrics is that of DIF. Keeping in mind the continuum of variation, conceptually, DIF is assessed by comparing the ICCs of different groups on an item. You can imagine that the same item is plotted separately for each group that the researcher wishes to evaluate (e.g., gender).

If the ICCs are identical for each group, or very close to identical, it can be said that the item does not display DIF. If, however, the ICCs are significantly different from one another across groups, then the item is said to show DIF. In most contexts, DIF is conceived of as a difference in placement (i.e., difficulty or threshold) of the two ICCS but as you will see in a few moments this does not necessarily have to be the case. Some examples of ICCs that do demonstrate DIF and some examples of items that do not demonstrate DIF will now be presented. Figure 3 is an example of an item that does not display DIF. As you can see, the area between the curves is very small and the parameters for each curve would be nearly equivalent.

Figure 3. An example of an item that does not display DIF.

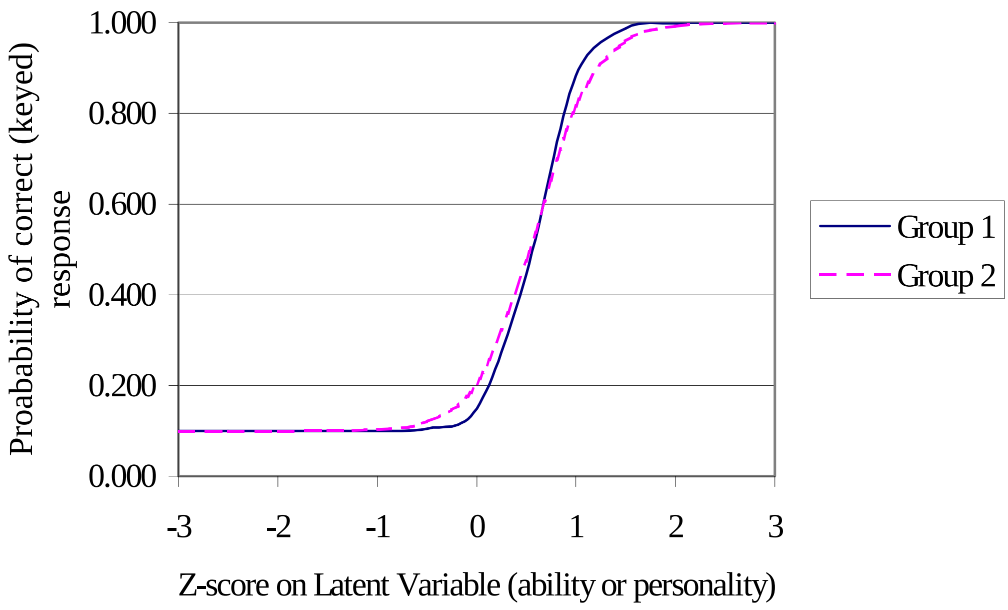


Figure 4, on the other hand, gives an example of an item that displays substantial DIF with a very large area between the two ICCs. This type of DIF is known as uniform DIF because the ICCs do not cross. An item such as the one shown in Figure 4 may not be an equivalent measure of the same latent variable for both groups.

Figure 4. An example of an item that displays substantial uniform DIF

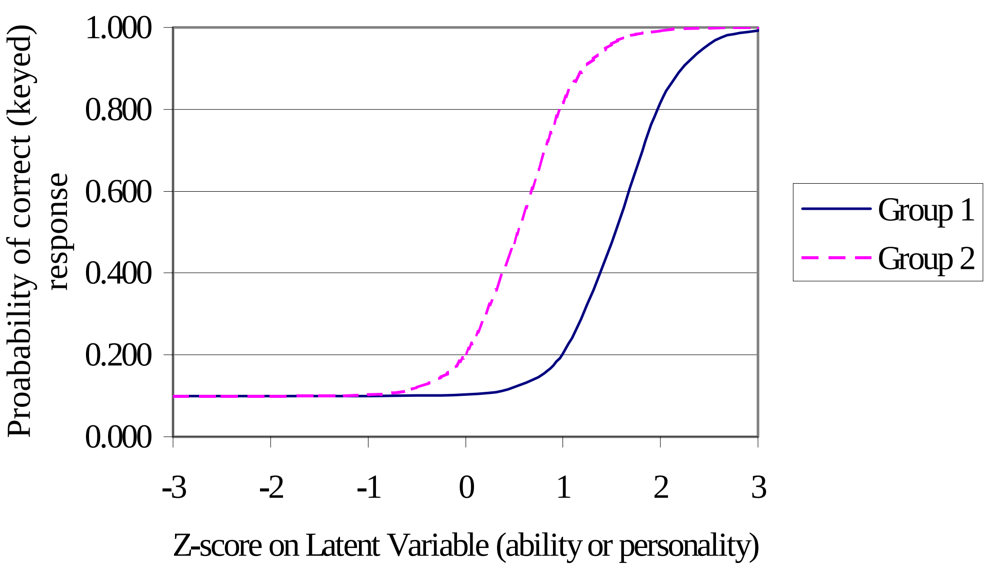
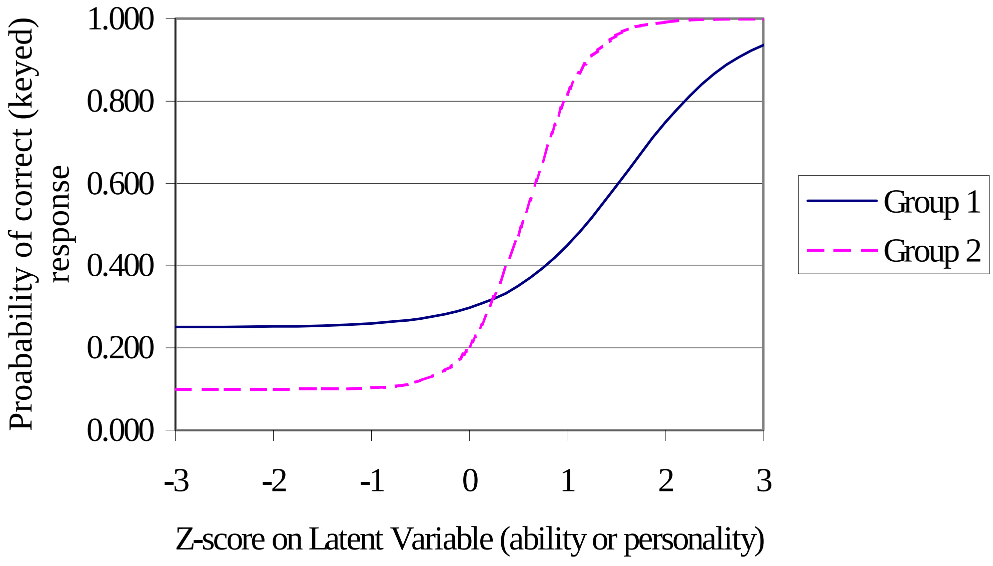


Figure 5 is an example of an item that displays substantial nonuniform DIF (i.e., the ICCs cross over one another). It depicts non-uniform DIF because for those individuals who score at or below the mean (i.e., z ≤ 0), Group 1 is favored whereas for those scoring above the mean (i.e., z>0) Group 2 is favored. It would be an understatement to say that the example in Figure 5 is a rather complex (and exaggerated) form of DIF; however, DIF which depends on where you score on the continuum (not to the degree depicted in Figure 5) does periodically arise in practice.

Figure 5. An example of an item that displays substantial nonuniform DIF



A Statistical Method For DIF: Logistic Regression

As was noted in the introduction, the two most commonly used scoring formats for tests and measures are binary and ordinal. Recall that it is not the question format that is important here but the scoring format. Items that are scored in a binary format are either: (a) items that are scored correct/incorrect in aptitude or achievement tests, or (b) items that are dichotomously scored according to a scoring key in personality or social measures. Items that are scored according to an ordinal scale may include Likert-type scales.

The issue of type of score becomes important in the DIF literature because historically DIF statistical methods have focussed on binary scored items. Most of the standard DIF methods are for binary items. For the case of ordinal responses, I will apply a statistical methodology that is a natural extension of the methodology for binary items. I will also introduce a new approach to measuring the magnitude of DIF for ordinal response formats. This new measure of DIF for ordinal variables is a natural extension of the strategy introduced by Zumbo and Thomas (1997) for binary item responses.

The purpose of this section is to provide a cursory introduction to logistic regression with enough detail to motivate the examples in the next section. Readers who plan to use logistic regression on a regular basis should consult Agresti (1996) for more details. I will first consider the statistical models, next the tests of significance for DIF, and finally the measures of magnitude of DIF (i.e., the effect sizes).

The Statistical Models

Binary Scored Items

For binary scored items, the detection of DIF can be accomplished through a number of different methods (see Clauser & Mazor, 1998; Green, 1994; Langenfeld, 1997). Currently, one of the most effective and recommended methods for detecting DIF is through the use of logistic regression (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990). Logistic regression is based on statistical modeling of the probability of responding correctly to an item by group membership (i.e., reference group and focal group - for example, non-minorities and visible minorities, respectively) and a criterion or conditioning variable. This criterion or conditioning variable is usually the scale or subscale total score but sometimes a different measure of the same variable.

The logistic regression procedure will use the item response (0 or 1) as the dependent variable, with grouping variable (dummy coded as 1=reference, 2=focal), total scale score for each subject (characterized as variable TOT) and a group by TOT interaction as independent variables. This method will provide a test of DIF conditionally on the relationship between the item response and the total scale score, testing the effects of group for uniform DIF, and the interaction of group and TOT to assess non-uniform DIF.

The logistic regression equation is

$$Y = b_0 + b_1\mathrm{TOT} + b_2\mathrm{GENDER} + b_3\mathrm{TOT}*\mathrm{GENDER}. \qquad (1)$$

where Y is a natural log of the odds ratio. That is, the equation

$$\ln\left[\frac{p_i}{(1-p_i)}\right] = b_0 + b_1 tot + b_2 group + b_3(tot * group), \qquad (2)$$

where *p* is the proportion of individuals that endorse the item in the direction of the latent variable. One can then test the 2-degree of freedom Chi-Square test for both uniform and non-uniform DIF.

Three advantages of using logistic regression (over other DIF methods such as the Mantel Haenszel) are that one:

- need not categorize a continuous criterion variable,
- can model uniform and/or non-uniform DIF (Swaminathan, 1994), and
- can generalize the binary logistic regression model for use with ordinal item scores.

Ordinal item scores

Ordinal logistic regression is but one method currently available for investigating DIF for items commonly found in personality and social psychological measures. I selected ordinal logistic regression rather than a generalized Mantel-Haenszel (M-H; Agresti, 1990) or *logistic discriminant* function analysis (Miller & Spray, 1993) because:

(a) using ordinal logistic regression has the advantage of using the same modeling strategy for binary and ordinal items,

(b) this common statistical model for binary and ordinal items should ease the process of implementation in an organization where DIF analyses are not yet common, and

(c) the Zumbo-Thomas DIF effect size method can be extended to ordinal logistic regression hence, unlike the other methods (e.g., generalized M-H or logistic discriminant function analysis), one has a test statistic *and* a natural corresponding measure of effect size.

Ordinal logistic regression has been previously proposed (e.g., Miller & Spray, 1993) for DIF of ordinal scored items but the development of the DIF effect size measure to ordinal items is presented for the first time in this handbook.

One can interpret logistic regression as a linear regression of predictor variables on an unobservable continuously distributed random variable, y^*. Thus, Equation (1) can be re-expressed as

$$\mathrm{y}^* = \mathrm{b}_0 + \mathrm{b}_1\mathrm{TOT} + \mathrm{b}_2\mathrm{GENDER} + \mathrm{b}_3\mathrm{TOT} * \mathrm{GENDER}_i + \varepsilon_i\ , \qquad (3)$$

where the ε_i are, for the logistic regression model, distributed with mean zero and variance $\pi^2\!/_3$. From this and some additional conditions, one can get an R^2 for ordinal logistic regression (see, Latila, 1993; McKelvey & Zavoina, 1975).

It is important to note that the notion of an unobservable continuously distributed random variable is quite familiar in the field of psychometrics. It is simply a latent continuum of variation. We normally conceive of unobserved variables as those which affect observable variables but which are not themselves observable because the observed magnitudes are subject to measurement error or because these variables do not correspond directly to anything that is likely to be measured. Examples are concepts such as "ability", "knowledge", or a personality variable such as "extroversion". Their existence is postulated and they are defined implicitly by the specification of the model and the methods used to estimate it.

Equation (3) reminds us that ordinal logistic regression assumes an unobservable continuum of variation (i.e., a latent variable)[2]. However, ordinal logistic regression can also be expressed as

$$\log\left[\frac{\mathrm{P(Y \le j)}}{\mathrm{P(Y > j)}}\right] = \alpha_j + b(\mathrm{X}), \text{ or}$$

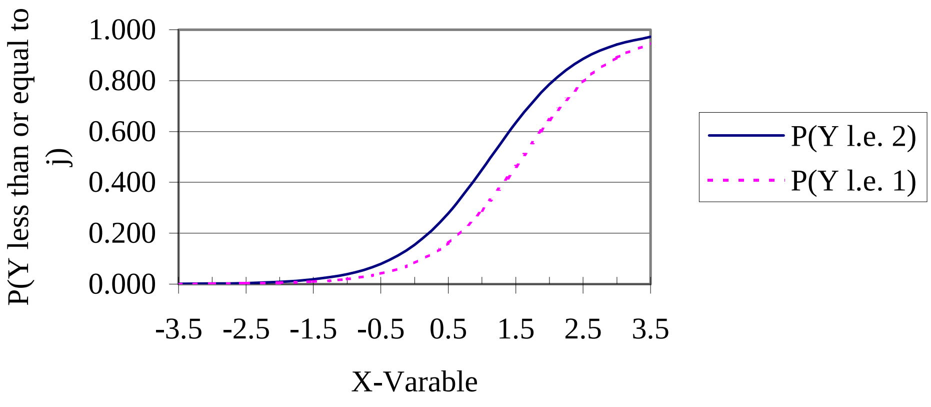
$$\mathrm{logit[P(Y \le j)]} = \alpha_j + b(\mathrm{X}), \tag{4}$$

where a logit is the natural logarithm of the ratio of two probabilities as seen in Equation (2), j = 1, 2, …, c-1, where c is the number of categories in the ordinal scale. The model requires a separate intercept parameter α_j for each cumulative probability. Equation (4) highlights two other assumptions of ordinal logistic regression over and above assumption of a continuum of variation (Agresti, 1996):

1. It operates on the principle of cumulative information along the latent variable. That is, for example, for a 3-point response an ordinal logistic regression model describes two relationships: the effect of X (in our case the total score for the scale) on the odds that $Y \le 1$ instead of $Y > 1$ on the scale, and the effect of X on the odds that $Y \le 2$ instead of $Y > 2$. Of course, for our three point scale, all of the responses will be less than or equal to three (the largest scale point) so it is not informative and hence left out of the model. The model requires two logistic curves (see Figure 6), one for each cumulative logit. The dashed line in Figure 6 depicts the cumulative probability for scoring less than or equal to 1 versus greater than 1; and the solid line depicts the cumulative probability for scoring less than or equal to 2 versus greater than 2.

[2] I am describing the cumulative logit model. Although I do not do so, one could also consider paired-category (e.g., adjacent-categories) logits for ordinal responses.

Figure 6. Curves for Cumulative Probabilities in a Cumulative Logit Model



2. At any given point on the X-axis the order of the two logistic curves is the same. For example, in Figure 6 at any point on the X-axis, if one starts at the X-axis going directly upward one will first come across the dashed line, where in the legend to the Figure P(Y le. 1) denotes "the probability that Y is less than or equal to one", and then the solid line, P(Y l.e. 2). This means that at any point on the X-axis the order of the lines are the same. This means that the logistic curves have a common slope, denoted *b* in Equation (4). When interpreting the magnitude of *b* please see Agresti (1996, pp. 213-214).

   For DIF analysis, Equation (4) can be written as

$$\text{logit}[\text{P}(\text{Y} \leq \text{j})] = \alpha_j + b_1 tot + b_2 group + b_3 (tot * group), \qquad (5)$$

In summary, if we had for example a 3-point Likert-type scale for an item, a ordinal logistic regression models the odds that someone will select the scale point 2 (or less) on the scale in comparison to selecting a response higher on the scale. Furthermore, the regression model does this for each point on the scale simultaneously. What a researchers ends up with is a regression equation having more than one intercept coefficient and only one slope. The common slope assumption could be tested with a nominal multinomial logit model.

Tests of Significance for DIF

Testing for the statistical significance of DIF follows naturally from its definition. That is, DIF modeling has a natural hierarchy of entering variables into the model. That is, Step #1: One first enters the conditioning variable (i.e., the total score),

Step #2: The group variable is entered, and finally

Step #3: The interaction term is entered into the equation.

With this information and the Chi-squared test for logistic regression one can compute the statistical tests for DIF. That is, one obtains the Chi-squared value for Step #3 and subtracts from it the Chi-squared value for Step #1. The resultant Chi-squared value can then be compared to its distribution function with 2 degrees of freedom. The 2 degrees of freedom arise from the fact that the model Chi-squared statistic at Step #3 is three and the model Chi-squared statistic at Step #1 is one (i.e., the difference in degrees of freedom is two). The resulting two-degree of freedom Chi-squared test is a simultaneous test of uniform and non-uniform DIF (Swaminathan & Rogers, 1990).

Swaminathan and Rogers (1990) focused their attention on the two degree-of-freedom Chi-squared test. Therefore, the corresponding effect size would be the R-squared attributable to both the group and interaction terms simultaneously (i.e., the R-squared at step #3 minus the R-squared at step #1). In logistic regression terminology, DIF is measurred by the simultaneous test of uniform *and* non-uniform DIF. However, the sequential modeling strategy described above allows one to also compare the R-squared values at step #2 to the R-squared value at step #1 to measure the unique variation attributable to the group differences over-and-above the conditioning variable (the total score) -- uniform DIF. Furthermore, comparing the R-squared values for step #3 and #2 measures the unique variation attributable to the interaction hence ascertaining how much of the DIF is non-uniform. This strategy needs to be tested further with real data but as our examples later demonstrate, it may be useful as a data analytic strategy.

The same three-step modeling strategy is used irrespective of whether one has binary or ordinal logistic regression. In its essence, this modeling strategy is akin to testing whether the group and interaction variables are statistically significant over-and-above the conditioning (i.e., matching) variable. As a note, a test of uniform DIF would have instead involved a difference between Steps #2 and #1. The advantage of logistic regression, of course, is that one can test for both uniform and non-uniform DIF simultaneously.

Measures of the Magnitude of DIF (Effect size)

Two points are noteworthy at this juncture. First, as per usual in statistical hypothesis testing, the test statistic should accompanied by some measure of the magnitude of the effect. This is necessary because small sample sizes can hide interesting statistical effects whereas large sample sizes (like the ones found in typical psychometric studies) can point to statistically significant findings where the effect is quite small and meaningless (Kirk, 1996). Second, I endorse the advice of Zumbo and Hubley (1998) who urge researchers to report effect sizes for both statistically significant and for statistically non-significant results. Following this practice, with time the psychometric community will have amassed an archive of effects for both statistically significant and non-significant DIF and therefore we can eventually move away from the somewhat arbitrary standards set by Cohen (1992).

Measuring the magnitude of DIF follows, as it should, the same strategy as the statistical hypothesis testing except that one only works with the R-squared values at each step. Zumbo and Thomas (1997) indicate that an examination of both the 2-df Chi-square test (of the likelihood ratio statistics) in logistic regression and a measure of effect size is needed to identify DIF. Without an examination of effect size, trivial effects could be statistically significant when the DIF test is based on a large sample size (i.e., too much statistical power). The Zumbo-Thomas measure of effect size for R^2 parallels effect size measures available for other statistics (see Cohen, 1992).

For an item to be classified as displaying DIF, the two-degree-of-freedom Chi-squared test in logistic regression had to have had a p-value less than or equal to 0.01 (set at this level because of the multiple hypotheses tested) *and* the Zumbo-Thomas effect size measure had to be at least an R-squared of 0.130. Pope (1997) has applied a similar criterion to binary personality items. It should be noted that Gierl and his colleagues (Gierl & McEwen, 1998, Gierl, Rogers, and Klinger, 1999) have adopted a more conservative criteria (i.e., the requisite R-squared for DIF is smaller) for the Zumbo-Thomas effect size in the context of educational measurement. They have also shown that the Zumbo-Thomas effect size measure is correlated with other DIF techniques like the Mantel-Haenszel and SIBTEST hence lending validity to the method.

In summary, I have found that a useful practice is to compute the R-squared effect for both (a) uniform DIF, and (b) a simultaneous test of uniform and non-uniform DIF. This strategy is useful because one is able to take advantage of the hierarchical nature of DIF modeling and therefore compare the R-squared for uniform DIF with the simultaneous uniform and non-uniform DIF to gage a sense of the magnitude or non-uniform DIF. The examples will demonstrate this approach.

Purifying the Matching Variable and Sample Size

You need to keep in mind that you should “purify the matching criterion” in the process of conducting the DIF analysis. That is, items that are identified as DIF are omitted, and the scale or total score is recalculated. This re-calculated total score is used as the matching criterion for a second logistic regression DIF analysis. Again, all items are assessed. This matching purification strategy has been shown to work empirically.

Holland and Thayer (1988) note that when the purification process is used, the item under examination should be included in the matching criterion even if it was identified as displaying DIF on initial screening and excluded from the criterion for all other items. That is, the item under study should always be included in its own matching criterion score. According to Holland and Thayer this reduces the number of Type I errors.

With regards to the matter of sample size, it has been shown in the literature that for binary items at least 200 people per group is adequate. However, the more people per group the better. Finally, it will aid the interpretation of the results if the sample does not have missing data. That is, one should conduct the analysis only on test takers from which you have complete data on the scale at hand (i.e., no missing data on the items comprising the scale and the grouping variable for your analysis).

Various R-squared Measures for DIF

Table 3 lists the various R-squared measures available to measure the magnitude of DIF. In short, the R-squared measures are used in the hierarchical sequential modeling process of adding more terms to the regression equation. The order of entering variables is determined by the definition of DIF.

Table 3.
R-squared Measures for DIF

Item Scoring	**Measure**	**Notes**
Ordinal	R-squared for ordinal	McKelvey & Zavoina (1975)
Binary (nominal)	Nagelkerke R-squared	Nagelkerke (c.f., Thomas & Zumbo, 1998)
Binary (nominal)	Weighted-least-squares R-squared	Thomas & Zumbo (1998)
Binary (ordinal)	R-squared for ordinal (i.e., same as above)	McKelvey & Zavoina (1975)

Unlike ordinal item scoring where there is one R-squared available, there are three options available for binary items: the Nagelkerke, weighted-least-squares (WLS), and ordinal R-squared measures.

Before reviewing the advantages of each of these methods, it is important to note that in line with conventional strategy in DIF analyses (Swaminathan & Rogers, 1990), the Nagelkerke and WLS treat the binary variable as nominal. However, one *can* treat the binary item scores from aptitude, achievement, knowledge, and personality/social measures as ordinal. Without loss of generality, let us imagine that the binary items are scored as "1" or "0" depending on whether the testee got the question correct or not, respectively, for knowledge or aptitude measures. Futhermore, one could score personality or social measures as "1" or "0" if the testee responding endorses the construct you are measuring or not, respectively. This is a traditional scoring rubric for binary items in aptitude, achievement, knowledge, or personality / social measures. In these cases, there is an implicit order of responses. That is, the responses scored as "1" are indicators for more knowledge, achievement, aptitude, or personality construct (depending on the setting) than a score of "0". For example, in a knowledge test, an item score of "1" is indicative that the respondent knows more than an individual who gets a score of "0". Therefore, it appears reasonable to treat these binary responses as, in essence, a larger ordinal response scale collapsed into two points.

Although each of the R-squared measures for binary items can be used in a hierarchical sequential modeling of DIF, each of these measures has its advantages[3]. That is,

(a) the Nagelkerke R-squared measure is easily obtained as output in a commonly used statistical package, SPSS;

(b) the WLS R-squared can be partitioned without resort to hierarchical sequential modeling (see Thomas & Zumbo, 1998; Zumbo & Thomas, 1997) and therefore is useful, for example, with multiple conditioning variables to order within blocks of variables;

(c) unlike the other two R-squared measures listed in Table 3, the ordinal logistic regression R-squared provides a uniform strategy for modeling ordered multicategory and binary items, and because it is assuming a latent continuous variable, it has the same properties as the R-squared we typically use in behavioral and social research involving continuous dependent variables.

To elaborate on the third point above, the ordinal logistic regression decomposes the variation in y^*, the latent continuous variable defined in Equation (3), into "explained" and "unexplained" components. As per the typical use of regression, this squared multiple correlation then represents the proportion of variation in the dependent variable captured by the regression and is defined as the regression sum of squares over the total sum of squares. Therefore, the R-squared values arising from the application of ordinal logistic regression are typical in magnitude to those found in behavioral and social science research and Cohen (1992) and Kirk's (1996) guidelines may be useful in interpretation.

Finally, although the ordinal logistic regression R-squared for measuring DIF in either multicategory scores and binary scores is introduced here, it is founded on the statistical theory for ordinal logistic regression, and the hierarchical sequential modeling strategy implicit in the definition of DIF. Future research applying this technique to a variety of measures will almost certainly result in a refinement of the approach and its interpretation.

[3] Note that not all R-squared measures for nominal binary responses can be used in a hierarchical regression.

Demonstrations With Examples

The various DIF analyses discussed in this handbook will be demonstrated with an example of ordinal scores and a second example of binary item scores. The data, and computer code for these examples, can be found at

http://quarles.unbc.ca/psyc/zumbo/DIF/index.html

I encourage you to access this website, take a copy of the programs and data sets to duplicate the results reported herein.

The Appendices list the SPSS syntax for the various DIF analyses listed in Table 3. Furthermore, some sample SPSS output is also listed in the Appendices to help the data analyst double-check their implementation of the methods. Appendices A through C list the SPSS syntax for the ordinal logistic regression, Nagelkerke, and WLS approaches, respectively.

It is important to recall that the SPSS syntax code assume complete data sets therefore you should use only complete data in your analyses.

Ordinal Item Scores

I will examine two items from a 20 item multicategory (Likert-type) questionnaire. Each question had a four-point scale ranging from 0 to 3. There are 511 subjects (249 females and 262 males). I will study each item for gender DIF.

Appendix A lists the SPSS syntax file (filename: running_ologit.sps) used to run the ordinal logistic regression. The data file used in this example is entitled "multicategory.sav". The SPSS file makes use of a public domain SPSS macro called ologit2.inc (written by By Prof. Dr. Steffen Kühnel, and modified by John Hendrickx University of Nijmegen The Netherlands).

Recall that there are 3 steps to DIF modeling (see page 26). For *item #1*, the Chi-squared value for the regression model with only the conditioning variable (step #1) is:

```
         LR-statistic
Chisqu.        DF    Prob.
216.153    1.000     .000
```

Adding both the uniform and non-uniform DIF terms to the model (at step #3 in the process) results in[4]:

```
         LR-statistic
Chisqu.        DF    Prob.
217.230    3.000     .000
```

The difference in Chi-squared values and degrees of freedom results in a 2-degree of freedom Chi-squared test of

Chi-squared : 1.077 with 2 d.f., p = 0.299

[4] The results of step #2 do not come into play until we turn to the effect sizes.

The Chi-squared and degree of freedom values were obtained by subtracting the results of the first model from the results from the model with both uniform and non-uniform DIF. The p-value was obtained from a standard statistics textbook or a computer program like EXCEL, Minitab, or Statistica. Given a non-significant p-value of 0.299, this item is not demonstrating DIF.

The effect size measures are as follows:

For the model with only the conditioning variable (total score) at step #1:

```
R-Square (%):
  40.82       or .4082
```

For the model with the conditioning and grouping variables at step #2:

```
R-Square (%):
  40.82       or .4082
```

For the model with the conditioning, grouping, and interaction variables (the model with both uniform and non-uniform DIF) at step #3:

```
R-Square (%):
  41.02        or .4102
```

Clearly, in accordance with the statistically non-significant 2 degree-of-freedom DIF test, adding the various variables does not increase the R-squared much at all. In fact, the DIF effect size for both uniform and non-uniform DIF (step #1 vs. step #3) is R-squared: 0.002 which by Cohen's (1992) criteria would correspond to a less than trivial effect size.

Having seen the detailed steps above for item #1, I will simply summarize the results for item #2. Example output for Item 2 can be found in Appendix D.

Step #1. Model with total score: χ^2(1)=111.181, R-squared=0.3135

Step #2. Uniform DIF: χ^2(2)=159.121, R-squared= 0.5010

Step #3. Uniform and Non-uniform DIF: χ^2(3)=161.194, R-squared= 0.5637

Examining the difference between steps #1 and #3 above we find a resulting χ^2(2)=50.013, p=0.00001, R-squared=0.2502 for the DIF test. Clearly, this item is statistically significant and shows a large DIF effect size. Using the critieria listed earlier in this handbook, item #2 clearly shows DIF. Moreover, comparing the R-squared values at steps #2 and #3, the data suggests that item #2 shows predominantly uniform DIF.

Binary Item Scores

The data file “binary.sav” contains simulated data from a 20 item test. Only items #1 and #2 are listed. Item #1 was simulated without DIF whereas item #2 was simulated with uniform DIF. There are 200 males and 200 females in this sample.

Table 4 lists the results from the DIF analysis of the two binary items. The R-squared for the 2 degrees-of-freedom DIF test can be found in the last column and was computed from the difference between R-squared values at Step #3 (fourth column) and Step #1 (second column). Furthermore, the R-squared at Step #2 can be compared to that at step #3 to see how much adding the non-uniform DIF variable contributes to the model.

Table 4.
Summary of the DIF analyses for the two binary items

	R-squared values at each step in the sequential hierarchical regression			**DIF $\chi^2(2)$ test**	**DIF R-squared**
Item #1	Step #1 Total score in the model	Step #2 Total score, and Uniform DIF variable in the model	Step #3 Total score, Uniform, and Non-uniform DIF variables in the model		
Nagelkerke	0.420	0.424	0.426	1.506, p=.4710	0.006
WLS	0.220	0.225	0.226	1.506, p=.4710	0.006
Ordinal	0.5625	0.5666	0.5867	2.505, p=.2858	0.024
Item #2					
Nagelkerke	0.158	0.345	0.345	69.32, p=.0000	0.187
WLS	0.208	0.697	0.700	69.32, p=.0000	0.402
Ordinal	0.156	0.3655	0.3677	69.06, p=.0000	0.210

Clearly, from the last column in Table 4, in all cases where there was statistically significant DIF the DIF R-squared values were substantially larger than where the Chi-squared test was statistically non-significant. Furthermore, for item #2 in Table 4, the difference in R-squared from Step #2 to Step #3 was quite small suggesting that the DIF was predominantly uniform.

As a final note, for the Nagelkerke and WLS approaches (unlike the ordinal approach) the two degree of freedom Chi-squared test is reported with its correct p-value in the output. For example, for item #2 I have highlighted the correct Chi-squared value and p-value.

```
Estimation terminated at iteration number 4 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      433.389
 Goodness of Fit        373.678
 Cox & Snell - R^2         .259
 Nagelkerke - R^2          .345

                    Chi-Square    df Significance

 Model                 119.687     3        .0000
 Block                  69.324     2        .0000
 Step                   69.324     2        .0000
```

In summary, this section presented examples of DIF analyses using ordinal and binary items. The accompanying SPSS syntax can be applied to the two data sets to reproduce the DIF results reported herein.

Concluding Remarks

In the tradition of logistic regression DIF tests, throughout this handbook, the term DIF is synonymous with the simultaneous test of uniform and non-uniform DIF with a 2 degree-of-freedom Chi-squared test. Moreover, I have highlighted the importance of reporting a measure of the corresponding effect size via R-squared.

Therefore, to conduct a DIF analysis you would:

(a) Perform the 3 step modeling as described on page 26,

(b) Compute the 2 degree-of-freedom Chi-squared test statistic,

(c) Compute the corresponding DIF R-squared for the 2 degree-of-freedom test,

(d) If after considering the Chi-squared test and the DIF effect size, you deem the DIF worthy of consideration by some criteria (such as the one we provide earlier in this handbook), you would study the R-squared values at each step of the 3 step modeling in (a) to gain insight into whether the DIF is predominantly uniform.

Clearly, (d) is a data analytic strategy used to help interpret the source of the DIF.

It should be noted that for measurement practitioners, DIF often means that there is some sort of systematic but construct irrelevant variance that is being tapped by the test or measure. Furthermore, the source for construct irrelevant variance is related to group membership. In a sense, this implies a multidimensionality whose presence and pervasiveness depends on group membership. The logistic regression modeling for investigating DIF with its corresponding measures of effect size is an essential starting point for bias analysis and in the long run will lead to more valid inferences from our tests and measures.

It appears to me, that for measurement practitioners, the uniform strategy of ordinal logistic regression I introduced for measuring the magnitude of DIF appears to be both

(a) a flexible modeling strategy that can handle both binary and ordinal multicategorical item scores, and

(b) a method that can be interpreted in the same manner as ordinary regression in other research contexts.

This uniform modeling method is grounded in the definition of DIF which emphasizes a sequential (hierarchical) model-building approach.

Bibliography

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123,* 207-215.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. (Vol. 4) Thousand Oaks, CA: Sage Publications.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Gierl, M. J., & McEwen, N. (1998). *Consistency among statistical methods and content review for identifying differential item functioning*. Presented at the *Measurement and Evaluation: Current and Future Research Directions for the New Millennium* conference, Banff, AB.

Gierl, M. J., Roger, W. T., & Klinger, D. (1999). *Using statistical and judgement reviews to identify and interpret translation DIF*. Presented at the Annual Meeting of the National Council for Measurement in Education (NCME), Montreal, Quebec.

Green, B. F. (1994). Differential item functioning: Techniques, findings, and prospects. In Dany Laveault, Bruno D. Zumbo, Marc E. Gessaroli, and Marvin W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues*. Ottawa, Canada: University of Ottawa.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations. *Educational Measurement: Issues and Practice, 16,* 20-26.

Latila, T. (1993). A pseudo-R^2 measure for limited and qualitative dependent variables. *Journal of Econometrics, 56,* 341-356.

McKelvey, R. D., & Zavoina, L (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology, 4,* 103-120.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

Pope, G. A. (1997). *Nonparametric item response modeling and gender differential item functioning of the Eysenck Personality Questionnaire.* Unpublished Master's thesis, University of Northern British Columbia, Prince George, B.C..

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Swaminathan, H. (1994). Differential item functioning: A discussion. In Dany Laveault, Bruno D. Zumbo, Marc E. Gessaroli, and Marvin W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues*. Ottawa, Canada: University of Ottawa.

Thomas, D. R., & Zumbo, B. D. (1998). *Variable importance in logistic regression based on partitioning an R-squared measure*. Presented at the Psychometric Society Meetings, Urbana, IL.

Zumbo, B. D. (Ed.) (1998). *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*. Netherlands: Kluwer Academic Press.

Zumbo, B . D., & Hubley, A. M. (1998 a). *Differential item functioning (DIF) analysis of a synthetic CFAT*. [Technical Note 98-4 , Personnel Research Team], Ottawa ON: Department of National Defense.

Zumbo, B. D., & Hubley, A. M. (1998 b). A note on misconceptions concerning prospective and retrospective power. *Journal of the Royal Statistical Society, Series D: The Statistician*, *47,* 385-388.

Zumbo, B. D., & Thomas, D. R. (1997) *A measure of effect size for a model-based approach for studying DIF.* Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.

Appendix A: SPSS Syntax for Ordinal Logistic Regression DIF

```
* SPSS SYNTAX written by:   .

* Bruno D. Zumbo, PhD   .
* Professor of Psychology and Mathematics,   .
* University of Northern British Columbia   .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the same .
*  folder .
* Change the filename, currently 'multicategory.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.

include file='ologit2.inc'.
execute.


GET
FILE='multicategory.sav'.
 EXECUTE .

compute item= item2.
compute total= total.
compute grp= group.

* Regression model with the conditioning variable, total score, in alone.
* Step #1 .
ologit var = item total
    /output=all.
execute.

* Regression model adding uniform DIF to model.
* Step #2.
ologit var = item total grp
    /contrast grp=indicator
    /output=all.
execute.


* Regression model adding non-uniform DIF to the model.
* Step #3.
ologit var = item total grp total*grp
    /contrast grp=indicator
    /output=all.
execute.
```

Appendix B: SPSS Syntax for Binary DIF with Nagelkerke R-squared

```
* SPSS SYNTAX written by:   .
* Bruno D. Zumbo, PhD   .
* Professor of Psychology and Mathematics,   .
* University of Northern British Columbia   .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.

GET
FILE='binary.sav'.
 EXECUTE .

compute item= item2.
compute total= scale.
compute grp= group.

* 2 df Chi-squared test and R-squared for the DIF (note that this is a simultaneous test .
* of  uniform and non-uniform DIF).

LOGISTIC REGRESSION VAR=item
  /METHOD=ENTER total  /METHOD=ENTER grp grp*total
  /CONTRAST (grp)=Indicator
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
execute.

*  1 df Chi-squared test and R-squared for uniform DIF.
*  This is particularly useful if one wants to determine the incremental R-squared .
*  attributed to the uniform DIF.

LOGISTIC REGRESSION VAR=item
  /METHOD=ENTER total  /METHOD=ENTER grp
  /CONTRAST (grp)=Indicator
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
execute.
```

Appendix C: SPSS Syntax for Binary DIF with WLS R-squared

```
* SPSS SYNTAX written by:   .
* Bruno D. Zumbo, PhD   .
* Professor of Psychology and Mathematics,   .
* University of Northern British Columbia   .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.

GET
FILE='binary.sav'.
 EXECUTE .

compute item= item1.
compute total= scale.
compute grp= group.

*    Aggregation.
*    Working with the Centered data.

*  Hierarchical regressions approach with the following order of steps:.
*      1. total.
*      2. total + group.
*      3. total + group + interac.
*  This also, of course, allows one to compute the relative Pratt Indices.


* Saves the standardized versions of group and total with the.
* eventual goal of centering before computing the cross-product term.

DESCRIPTIVES
  VARIABLES=group total  /SAVE
  /FORMAT=LABELS NOINDEX
  /STATISTICS=MEAN STDDEV MIN MAX
  /SORT=MEAN (A) .

* Allows for both uniform and non-uniform DIF.
* Provides the 2df Chi-square test for DIF.

LOGISTIC REGRESSION item
  /METHOD=ENTER ztotal /method=enter zgroup ztotal*zgroup
  /SAVE PRED(pre1).
execute.
```

```
*  The following command is required to deal with the repeaters in.
*  the data.  The WLS regression will be conducted on the aggregate.
*  file entitled "AGGR.SAV".

AGGREGATE
  /OUTFILE='aggr.sav'
  /BREAK=zgroup ztotal
  /item = SUM(item) /pre1 = MEAN(pre1)
  /Ni=N.

GET
  FILE='aggr.sav'.
EXECUTE .

compute interact=zgroup*ztotal.
execute.

COMPUTE v1 = Ni*pre1 *(1 - pre1) .
EXECUTE .

COMPUTE z1 = LN(pre1/(1-pre1))+ (item-Ni*pre1)/Ni/pre1/(1-pre1) .
EXECUTE .

FORMATS v1, z1 (F8.4).
execute.

* Overall logistic regression.
* Both Uniform and Non-uniform DIF.
REGRESSION
  /MISSING LISTWISE
  /REGWGT=v1
  /descriptives=corr
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHA
  /NOORIGIN
  /DEPENDENT z1
  /METHOD=ENTER ztotal / method=enter zgroup / method= enter interact .
execute.
```

Appendix D: Output for Item 2 of Ordinal Item Score

```
* SPSS SYNTAX written by:   .
* Bruno D. Zumbo, PhD   .
* Professor of Psychology and Mathematics,   .
* University of Northern British Columbia   .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Copy this file and the file "ologit2.inc", and your SPSS data file into the sa
   me folder .
* Change the filename, currently 'multicategory.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file
   .
* Run this entire syntax command file.

include file='ologit2.inc'.
4835  set printback off.

>Warning # 235
>The position and length given in a macro SUBSTR function are inconsistent
>with the string argument.  The null string has been used for the result.

Matrix
>Error # 12302 on line 4851 in column 256.  Text: (End of Command)
>Syntax error.
>This command not executed.

>Error # 12548 on line 4852 in column 16.  Text:
>MATRIX syntax error: unrecognized or not allowed token encountered.
>This command not executed.

Scan error detected in parser.

------ END MATRIX -----
>Note # 213
>Due to an error, INCLUDE file processing has been terminated.  All
>transformations since the last procedure command have been discarded.


>Note # 236
>All outstanding macros have been terminated, all include processing has
>been terminated, and all outstanding PRESERVE commands have been undone.

execute.

GET
FILE='multicategory.sav'.
 EXECUTE .

compute item= item2.
compute total= total.
compute grp= group.

* Regression model with the conditioning variable, total score, in alone.
ologit var = item total
     /output=all.
Matrix
Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

                 (by Steffen M. KUEHNEL)

******************** Information Section ********************
Dependent variable is:
 item

Marginal distribution of dependent variable
     Value    Frequ.   Percent   %>Value
       .00    435.00     85.13     14.87
      1.00     52.00     10.18      4.70
      2.00     18.00      3.52      1.17
      3.00      6.00      1.17       .00

Effective sample size:    511
```

```
Means and standard deviations of independent variables:
            Mean   Std.Dev.
total     9.9413     9.2916

******************** Estimation Section ********************

Running Iteration No.:
   1

Running Iteration No.:
   2

Running Iteration No.:
   3

Running Iteration No.:
   4

Running Iteration No.:
   5
..... Optimal solution found.

******************** OUTPUT SECTION ********************
LR-test that all predictor weights are zero
-------------------------------------------

-2 Log-Likelihood of Model with Constants only:
   551.535

-2 Log-Likelihood of full Model:
   440.354

LR-statistic
  Chisqu.       DF    Prob. %-Reduct
  111.181    1.000     .000     .202

Estimations, standard errors, and effects
-----------------------------------------

            Coeff.=B     Std.Err.     B/Std.E.        Prob.       exp(B)
total        .131920      .013416     9.833118      .000000     1.141017
Const.1    -3.463503      .255288   -13.567024      .000000      .031320
Const.2    -5.079443      .352167   -14.423409      .000000      .006223
Const.3    -6.822095      .548289   -12.442512      .000000      .001089

Results assuming a latent continuous variable
---------------------------------------------

R-Square (%):
  31.35

Standardized regression weights of the latent variable:
total   .5599

------ END MATRIX -----
execute.

* Regression model adding uniform DIF to model.
ologit var = item total grp
     /contrast grp=indicator
     /output=all.
Matrix
Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE

                 (by Steffen M. KUEHNEL)

Parameter coding for grp using the indicator contrast
 Value  Freq grp.1
     1   249     1
     2   262     0
```

```
******************** Information Section ********************


Dependent variable is:
 item

Marginal distribution of dependent variable
     Value    Frequ.   Percent   %>Value
       .00    435.00     85.13     14.87
      1.00     52.00     10.18      4.70
      2.00     18.00      3.52      1.17
      3.00      6.00      1.17       .00

Effective sample size:
    511

Means and standard deviations of independent variables:
            Mean   Std.Dev.
total     9.9413     9.2916
grp.1      .4873      .5003
******************** Estimation Section ********************

Running Iteration No.:
   1

Running Iteration No.:
   2

Running Iteration No.:
   3

Running Iteration No.:
   4

Running Iteration No.:
   5

Running Iteration No.:
   6
..... Optimal solution found.

******************** OUTPUT SECTION ********************

LR-test that all predictor weights are zero
-------------------------------------------

-2 Log-Likelihood of Model with Constants only:
   551.535

-2 Log-Likelihood of full Model:
   392.414

LR-statistic
  Chisqu.       DF    Prob. %-Reduct
  159.121    2.000     .000     .289

Estimations, standard errors, and effects
-----------------------------------------
            Coeff.=B     Std.Err.     B/Std.E.        Prob.       exp(B)
total        .144826      .014781     9.798088      .000000     1.155839
grp.1       2.268576      .383818     5.910550      .000000     9.665630
Const.1    -5.188266      .465705   -11.140668      .000000      .005582
Const.2    -6.926669      .545849   -12.689722      .000000      .000981
Const.3    -8.668915      .689934   -12.564841      .000000      .000172
Results assuming a latent continuous variable
---------------------------------------------
R-Square (%):
  50.10

Standardized regression weights of the latent variable:
total   .5241
grp.1   .4420

------ END MATRIX -----
execute.
```

```
* Regression model adding non-uniform DIF to the model.
ologit var = item total grp total*grp
     /contrast grp=indicator
     /output=all.
```

Matrix

```
Run MATRIX procedure:

LOGISTIC REGRESSION with an ORDINAL DEPENDENT VARIBLE
           (by Steffen M. KUEHNEL)

Parameter coding for grp using the indicator contrast
 Value  Freq grp.1
     1   249     1
     2   262     0

Interaction term total*grp
 int1.1   total    grp.1
******************** Information Section ********************
Dependent variable is:
 item

Marginal distribution of dependent variable
     Value    Frequ.   Percent   %>Value
       .00    435.00     85.13     14.87
      1.00     52.00     10.18      4.70
      2.00     18.00      3.52      1.17
      3.00      6.00      1.17       .00

Effective sample size:
    511

Means and standard deviations of independent variables:
             Mean   Std.Dev.
total      9.9413     9.2916
grp.1       .4873      .5003
int1.1     5.1546     8.3997

******************** Estimation Section ********************

Running Iteration No.:
   1

Running Iteration No.:
   2

Running Iteration No.:
   3

Running Iteration No.:
   4

Running Iteration No.:
   5

Running Iteration No.:
   6

..... Optimal solution found.

******************** OUTPUT SECTION ********************

LR-test that all predictor weights are zero
-------------------------------------------

-2 Log-Likelihood of Model with Constants only:
   551.535

-2 Log-Likelihood of full Model:
   390.341

LR-statistic
  Chisqu.       DF    Prob. %-Reduct
  161.194    3.000     .000     .292
```

```
Estimations, standard errors, and effects
-----------------------------------------


            Coeff.=B     Std.Err.     B/Std.E.        Prob.       exp(B)
total        .174374      .026730     6.523460      .000000     1.190500
grp.1       3.212986      .826100     3.889340      .000101    24.853177
int1.1      -.043721      .031132    -1.404351      .160214      .957221
Const.1    -5.920075      .773599    -7.652644      .000000      .002685
Const.2    -7.650652      .823231    -9.293440      .000000      .000476
Const.3    -9.404898      .937163   -10.035501      .000000      .000082

Results assuming a latent continuous variable
---------------------------------------------

R-Square (%):
  56.37

Standardized regression weights of the latent variable:
total    .5901
grp.1    .5854
int1.1  -.1337

------ END MATRIX -----
execute.
```

APPENDIX E: Output for Nagelkerke R-squared

```
* SPSS SYNTAX written by:   .
* Bruno D. Zumbo, PhD   .
* Professor of Psychology and Mathematics,   .
* University of Northern British Columbia   .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file
   .
* Run this entire syntax command file.

GET
FILE='binary.sav'.
 EXECUTE .

compute item= item2.
compute total= scale.
compute grp= group.

*  2 df Chi-squared test and R-squared for the DIF (note that this is a simultan
   eous test of uniform .
*  and non-uniform DIF).

LOGISTIC REGRESSION VAR=item
  /METHOD=ENTER total  /METHOD=ENTER grp grp*total
  /CONTRAST (grp)=Indicator
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Logistic Regression

```
_


      Total number of cases:      400 (Unweighted)
      Number of selected cases:   400
      Number of unselected cases: 0

      Number of selected cases:                 400
      Number rejected because of missing data:  0
      Number of cases included in the analysis: 400



Dependent Variable Encoding:

Original       Internal
Value          Value
     .00       0
    1.00       1
_


                               Parameter
                 Value   Freq  Coding
                                 (1)
GRP
                  1.00    200  1.000
                  2.00    200   .000


      Interactions:

INT_1    GRP(1) by TOTAL
_Dependent Variable..   ITEM

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   553.07688

* Constant is included in the model.

Beginning Block Number  1.  Method: Enter
```

```
Variable(s) Entered on Step Number
1..       TOTAL

Estimation terminated at iteration number 3 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      502.714
 Goodness of Fit        393.328
 Cox & Snell - R^2         .118
 Nagelkerke - R^2          .158

                     Chi-Square    df Significance

 Model                   50.363     1        .0000
 Block                   50.363     1        .0000
 Step                    50.363     1        .0000

Classification Table for ITEM
The Cut Value is .50
                    Predicted
                   .00    1.00     Percent Correct
                     0  I    1
Observed        +-------+-------+
   .00      0   I  159  I   53  I   75.00%
                +-------+-------+
   1.00     1   I   88  I  100  I   53.19%
                +-------+-------+
                           Overall  64.75%

---------------------- Variables in the Equation -----------------------

Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

TOTAL          .1959     .0302  42.2113     1    .0000   .2696    1.2164
Constant     -2.1563     .3332  41.8696     1    .0000


Beginning Block Number  2.  Method: Enter

Variable(s) Entered on Step Number
1..       GRP
          GRP * TOTAL
_


Estimation terminated at iteration number 4 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      433.389
 Goodness of Fit        373.678
 Cox & Snell - R^2         .259
 Nagelkerke - R^2          .345

                     Chi-Square    df Significance

 Model                  119.687     3        .0000
 Block                   69.324     2        .0000
 Step                    69.324     2        .0000

Classification Table for ITEM
The Cut Value is .50
                    Predicted
                   .00    1.00     Percent Correct
                     0  I    1
Observed        +-------+-------+
   .00      0   I  161  I   51  I   75.94%
                +-------+-------+
   1.00     1   I   65  I  123  I   65.43%
                +-------+-------+
                           Overall  71.00%

---------------------- Variables in the Equation -----------------------

Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

TOTAL          .2951     .0518  32.5063     1    .0000   .2463    1.3433
```

```
GRP(1)       -2.4802     .8723   8.0833     1    .0045  -.1100     .0837
INT_1          .0373     .0772    .2342     1    .6284   .0000    1.0380
Constant     -2.1928     .4719  21.5942     1    .0000
execute.

*  1 df Chi-squared test and R-squared for uniform DIF.
*  This is particularly useful if one wants to determine the incremental R-squar
   ed .
*  attributed to the uniform DIF.

LOGISTIC REGRESSION VAR=item
  /METHOD=ENTER total  /METHOD=ENTER grp
  /CONTRAST (grp)=Indicator
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
Logistic Regression
_


      Total number of cases:      400 (Unweighted)
      Number of selected cases:   400
      Number of unselected cases: 0

      Number of selected cases:                 400
      Number rejected because of missing data:  0
      Number of cases included in the analysis: 400



Dependent Variable Encoding:

Original       Internal
Value          Value
     .00       0
    1.00       1

                               Parameter
                 Value   Freq  Coding
                                 (1)
GRP
                  1.00    200  1.000
                  2.00    200   .000
Dependent Variable..   ITEM

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   553.07688

* Constant is included in the model.


Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..       TOTAL

Estimation terminated at iteration number 3 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      502.714
 Goodness of Fit        393.328
 Cox & Snell - R^2         .118
 Nagelkerke - R^2          .158

                     Chi-Square    df Significance

 Model                   50.363     1        .0000
 Block                   50.363     1        .0000
 Step                    50.363     1        .0000
```

```
Classification Table for ITEM
The Cut Value is .50
                    Predicted
                   .00    1.00     Percent Correct
                    0  I    1
Observed        +-------+-------+
   .00      0   I  159  I   53  I   75.00%
                +-------+-------+
   1.00     1   I   88  I  100  I   53.19%
                +-------+-------+
                           Overall  64.75%

---------------------- Variables in the Equation -----------------------

Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

TOTAL          .1959     .0302  42.2113     1    .0000   .2696    1.2164
Constant     -2.1563     .3332  41.8696     1    .0000


Beginning Block Number  2.  Method: Enter

Variable(s) Entered on Step Number
1..       GRP
_


Estimation terminated at iteration number 4 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      433.624
 Goodness of Fit        374.540
 Cox & Snell - R^2         .258
 Nagelkerke - R^2          .345

                     Chi-Square    df Significance

 Model                  119.452     2        .0000
 Block                   69.089     1        .0000
 Step                    69.089     1        .0000

Classification Table for ITEM
The Cut Value is .50
                    Predicted
                   .00    1.00     Percent Correct
                    0  I    1
Observed        +-------+-------+
   .00      0   I  161  I   51  I   75.94%
                +-------+-------+
   1.00     1   I   65  I  123  I   65.43%
                +-------+-------+
                           Overall  71.00%

---------------------- Variables in the Equation -----------------------

Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

TOTAL          .3125     .0384  66.2444     1    .0000   .3575    1.3669
GRP(1)       -2.0828     .2780  56.1269     1    .0000  -.3281     .1246
Constant     -2.3415     .3667  40.7653     1    .0000
execute.
```

Appendix F: Output for the WLS R-squared DIF

```
* SPSS SYNTAX written by:   .
* Bruno D. Zumbo, PhD   .
* Professor of Psychology and Mathematics,   .
* University of Northern British Columbia   .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file
   .
* Run this entire syntax command file.

GET
FILE='binary.sav'.
 EXECUTE .

compute item= item2.
compute total= scale.
compute grp= group.

*    Aggregation.
*    Working with the Centered data.

*  Hierarchical regressions approach with the following order of steps:.
*      1. total.
*      2. total + group.
*      3. total + group + interac.
*  This also, of course, allows one to compute the relative Pratt Indices.


* Saves the standardized versions of group and total with the.
* eventual goal of centering before computing the cross-product term.

DESCRIPTIVES
  VARIABLES=group total  /SAVE
  /FORMAT=LABELS NOINDEX
  /STATISTICS=MEAN STDDEV MIN MAX
  /SORT=MEAN (A) .
```

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
GROUP	400	1.00	2.00	1.5000	.5006
TOTAL	400	.00	20.00	10.3050	3.9858
Valid N (listwise)	400				

```
* Allows for both uniform and non-uniform DIF.
* Provides the 2df Chi-square test for DIF.

LOGISTIC REGRESSION item
  /METHOD=ENTER ztotal /method=enter zgroup ztotal*zgroup
  /SAVE PRED(pre1).
```

Logistic Regression

```
_


      Total number of cases:      400 (Unweighted)
      Number of selected cases:   400
      Number of unselected cases: 0

      Number of selected cases:                 400
      Number rejected because of missing data:  0
      Number of cases included in the analysis: 400



Dependent Variable Encoding:

Original       Internal
Value          Value
     .00       0
    1.00       1


      Interactions:

INT_1    ZTOTAL by ZGROUP
_


Dependent Variable..   ITEM

Beginning Block Number  0.  Initial Log Likelihood Function
```

```
-2 Log Likelihood   553.07688

* Constant is included in the model.


Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..       ZTOTAL    Zscore(TOTAL)

Estimation terminated at iteration number 3 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      502.714
 Goodness of Fit        393.328
 Cox & Snell - R^2         .118
 Nagelkerke - R^2          .158

                     Chi-Square    df Significance

 Model                   50.363     1        .0000
 Block                   50.363     1        .0000
 Step                    50.363     1        .0000

Classification Table for ITEM
The Cut Value is .50
                    Predicted
                   .00    1.00     Percent Correct
                     0  I    1
Observed        +-------+-------+
   .00      0   I  159  I   53  I   75.00%
                +-------+-------+
   1.00     1   I   88  I  100  I   53.19%
                +-------+-------+
                           Overall  64.75%

---------------------- Variables in the Equation -----------------------
Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

ZTOTAL         .7809     .1202  42.2113     1    .0000   .2696    2.1834
Constant      -.1374     .1069   1.6537     1    .1985


Beginning Block Number  2.  Method: Enter

Variable(s) Entered on Step Number
1..       ZGROUP    Zscore(GROUP)
          ZTOTAL * ZGROUP
_
```

```
Estimation terminated at iteration number 4 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood      433.389
 Goodness of Fit        373.678
 Cox & Snell - R^2         .259
 Nagelkerke - R^2          .345

                     Chi-Square    df Significance

 Model                  119.687     3        .0000
 Block                   69.324     2        .0000
 Step                    69.324     2        .0000

Classification Table for ITEM
The Cut Value is .50
                    Predicted
                   .00    1.00     Percent Correct
                     0  I    1
Observed        +-------+-------+
   .00      0   I  161  I   51  I   75.94%
                +-------+-------+
   1.00     1   I   65  I  123  I   65.43%
                +-------+-------+
                           Overall  71.00%

---------------------- Variables in the Equation -----------------------

Variable           B      S.E.     Wald    df      Sig       R    Exp(B)

ZTOTAL        1.2507     .1538  66.1742     1    .0000   .3573    3.4930
ZGROUP        1.0490     .1403  55.9035     1    .0000   .3275    2.8548
INT_1         -.0745     .1539    .2342     1    .6284   .0000     .9282
Constant      -.1991     .1401   2.0195     1    .1553

1 new variables have been created.
  Name         Contents

  PRE1         Predicted Value
execute.

*  The following command is required to deal with the repeaters in.
*  the data.  The WLS regression will be conducted on the aggregate.
*  file entitled "AGGR.SAV".

AGGREGATE
  /OUTFILE='aggr.sav'
  /BREAK=zgroup ztotal
  /item = SUM(item) /pre1 = MEAN(pre1)
  /Ni=N.

GET
  FILE='aggr.sav'.
EXECUTE .

compute interact=zgroup*ztotal.
execute.

COMPUTE v1 = Ni*pre1 *(1 - pre1) .
EXECUTE .

COMPUTE z1 = LN(pre1/(1-pre1))+ (item-Ni*pre1)/Ni/pre1/(1-pre1) .
EXECUTE .

FORMATS v1, z1 (F8.4).
execute.

* Overall logistic regression.
* Both Uniform and Non-uniform DIF.
REGRESSION
  /MISSING LISTWISE
  /REGWGT=v1
  /descriptives=corr
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHA
  /NOORIGIN
  /DEPENDENT z1
  /METHOD=ENTER ztotal / method=enter zgroup / method= enter interact .
```

Regression

Correlations[a]

		Z1	Zscore(TOTAL)	Zscore(GROUP)	INTERACT
Pearson Correlation	Z1	1.000	.456	.343	.055
	Zscore(TOTAL)	.456	1.000	-.540	.046
	Zscore(GROUP)	.343	-.540	1.000	.070
	INTERACT	.055	.046	.070	1.000

a. Weighted Least Squares Regression - Weighted by V1

Variables Entered/Removed[b,c]

Model	Variables Entered	Variables Removed	Method
1	Zscore(TOTAL)[a]	.	Enter
2	Zscore(GROUP)[a]	.	Enter
3	INTERACT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Z1

c. Weighted Least Squares Regression - Weighted by V1

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.456[a]	.208	.186	1.583467	.208	9.435	1	36	.004
2	.835[b]	.697	.680	.992311	.490	56.670	1	35	.000
3	.836[c]	.700	.673	1.003371	.002	.233	1	34	.633

a. Predictors: (Constant), Zscore(TOTAL)

b. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP), INTERACT

ANOVA[d,e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.656	1	23.656	9.435	.004[a]
	Residual	90.265	36	2.507		
	Total	113.921	37			
2	Regression	79.457	2	39.729	40.347	.000[b]
	Residual	34.464	35	.985		
	Total	113.921	37			
3	Regression	79.692	3	26.564	26.386	.000[c]
	Residual	34.230	34	1.007		
	Total	113.921	37			

a. Predictors: (Constant), Zscore(TOTAL)

b. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP), INTERACT

d. Dependent Variable: Z1

e. Weighted Least Squares Regression - Weighted by V1

Coefficients[a,b]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-9.55E-02	.185		-.515	.610		
	Zscore(TOTAL)	.626	.204	.456	3.072	.004	1.000	1.000
2	(Constant)	-.162	.117		-1.391	.173		
	Zscore(TOTAL)	1.243	.152	.905	8.190	.000	.709	1.411
	Zscore(GROUP)	1.041	.138	.831	7.528	.000	.709	1.411
3	(Constant)	-.199	.141		-1.416	.166		
	Zscore(TOTAL)	1.251	.154	.910	8.107	.000	.701	1.426
	Zscore(GROUP)	1.049	.141	.838	7.452	.000	.699	1.430
	INTERACT	-7.45E-02	.154	-.046	-.482	.633	.985	1.015

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by V1

```
execute.
```